Testing combinatorial correlation by decision diagrams

06 そのデータ、本当に偏ってますか？

決定グラフを用いた組合せ的相関検定

どんな研究

病気が流行したとき、その流行に地域性があるかないかを知ることは重要です。組合せ的相関検定では観測データである地域別病気発症率と地域の隣接関係から流行の地域性の有無を検定します。具体的には観測データの偏りとその稀さを計算し、病気の地域性の有無を判定します。

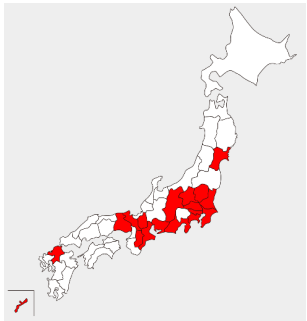
どこが凄い

偏りはScan統計量として、その稀さはP値（その統計量が考えられる全ての観測パターンのうち上位何%であるか）として計算されます。これらを愚直に計算すると47都道府県別の検定には1億年以上かかります。私達はこの検定時間を決定グラフと呼ばれるデータ構造を用いることで1日に短縮します。

めざす未来

本手法は流行の地域性のみでなく、センサ網上の侵入者の検知や購買履歴中の顕著な組合せの発見など、様々な組合せ構造に適用可能な一般的な手法です。本手法により、日々蓄積される様々な組合せデータに潜む重要な情報を自動的に抽出可能になります。

組合せ的相関検定　「観測は構造に依存している？」に回答



（例1）病気の流行

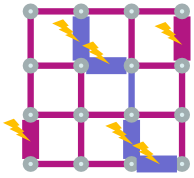
赤: 病気が流行している都道府県

白: 病気が流行していない都道府県

※この観測は説明のための例です。

この病気の流行は

1. 地域性がある？　（帰無仮説）
2. 偶然偏っている？（対立仮説）



（例2）センサ網

センサの反応 ⚡ は

1. 侵入者？
2. ノイズ？

購買履歴	頻度
A B C	3
A B	5
B C	1
A	1
B	1

（例3）購買履歴

商品 A B の組合せは

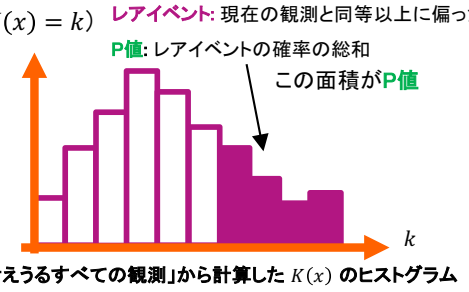
1. 人気？
2. 偶然？

この問題の難しさ　観測の統計量が「考えられる全ての観測」のうち上位何%であるかの計算が必要

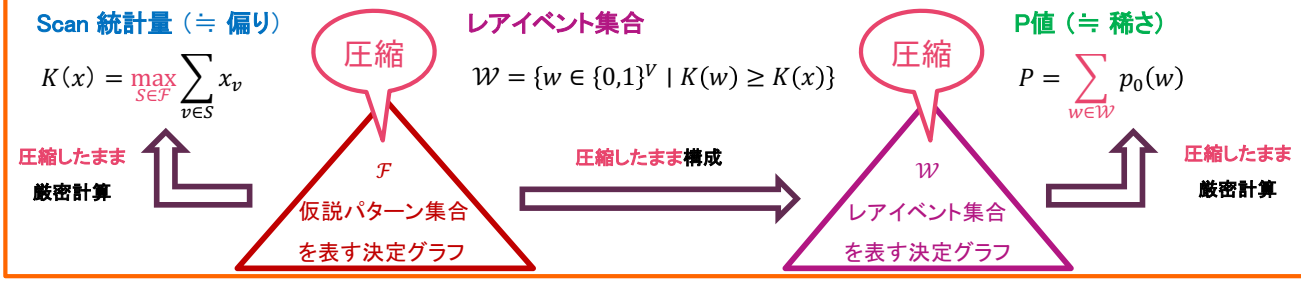
組合せ的相関検定の手順

1. 観測 x と 仮説パターン集合 $\mathcal{F}$（例：全隣接都道府県の集合）を受け取る
2. 観測 x の Scan 統計量（≒ 偏り）$K(x)$ を計算
3. 統計量 $K(x)$ の P値（≒ 稀さ）を レアイベント集合 $\mathcal{W}$ から計算
4. P値 ≦ 有意水準（例 0.05）なら対立仮説を棄却 ⇒ **地域性有り**

$\mathcal{F}$ と $\mathcal{W}$ は 指数的に巨大な集合 ➡ 愚直に計算すると（例1）で1億年以上かかる 以下の提案技術を用いれば1日で計算可能！



提案技術　仮説パターン集合とレアイベント集合を決定グラフで圧縮してP値を効率的に計算



関連文献

[1] M. Ishihata, T. Maehara, "Exact Bernoulli scan statistics by binary decision diagrams," *The 28th International Joint Conference on Artificial Intelligence (IJCAI-2019)*, 2019.

連 絡 先

石畠 正和（Masakazu Ishihata） 協創情報研究部 知能創発環境研究グループ

Email: cs-openhouse-ml@hco.ntt.co.jp

オープンハウス 2020